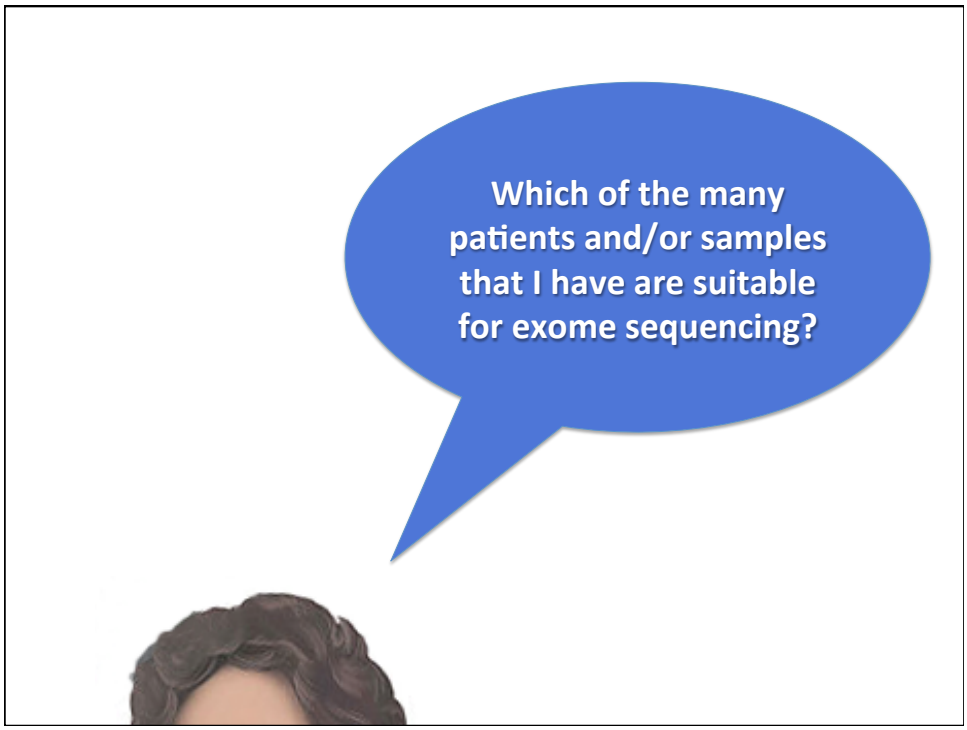
Exome 101: Filtering strategies for identifying germline variants that cause disease

Leslie G. Biesecker, M.D.
Genetic Disease Research Branch
National Human Genome Research Institute



Good & bad news

- Can work wonders
  - Small families
  - Stuck positional cloning projects
  - *de novo* dominants
  - Others
- May only work 30-40% of the time
  - Publication bias
  - Many biological & technical reasons
  - Gene ID not adequate for good paper

General outline

- What is in an exome (and what is not)
- The differences of exome sequencing vs. positional cloning
- How to do it
  - Example of X-linked
  - Example of recessive
  - Example of dominant
  - Example of sporadic *de novo*
  - Example of mosaic

What is a 'whole' exome sequence?

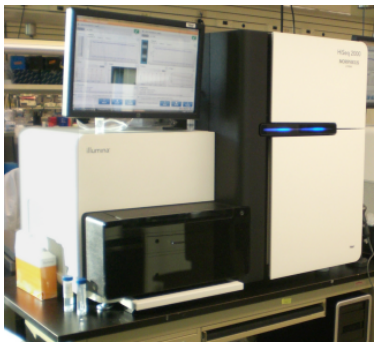
- The sequence of all exons of the genome
  - Not all genes are recognized
  - Not all exons of recognized genes are known
  - Non-coding exons not always targeted
  - Not all targeted exons are well-captured
  - Not all targeted sequences can be aligned
  - Not all aligned sequences can be accurately called

  - Not all that 'whole…'

What is missing from a WES?

- Some genes
- Some parts of some genes
- Non-genic control elements
- Non-canonical splice elements
- Structural DNA assessments
- CNVs
- mtDNA
- Some miRNAs

- If your disease is caused by one of these, WES is the wrong approach

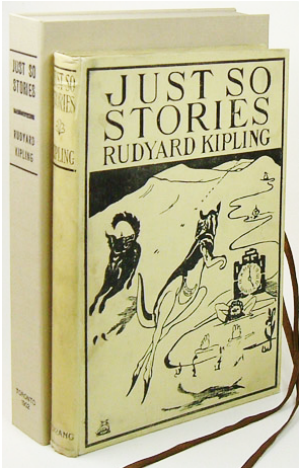


WES vs. Positional cloning

- WES
  - Small families OK
  - Locus homogeneity is very important
    - Hard to fix post WES
  - Allelic heterogeneity is very important
  - Bird in hand vs…
  - Phenocopies not a big issue (usually)
- Positional cloning
  - Large families essential
  - Locus heterogeneity not a big issue
    - Easy to assess @ linkage
  - Allelic heterogeneity not a big issue
  - Hammer candidates
  - Phenocopies a big issue for meiotic mapping

WES vs. Positional cloning

- Absence of genetic mapping is disadvantage
- >20,000 candidates
  - Chance of Type I error is high
  - Without meiotic mapping you will need additional sources of evidence for causation



X-linked disorder: TARP

- X-linked 'recessive'
- Cleft palate, heart defects, club feet
- Severe - 100% male lethality
- Ultra-rare (two known families)
- Little DNA on boys, sequenced carriers

X-linked disorder: TARP

- X 'exome' capture
  - Region: 2,675,000 – 154,500,000 bp
  - All UCSC coding exons
  - Reads: 20,262,045; 18,775,942
  - Sequence: 729,433,620; 675,933,912 bp
  - Aligned to X exome: 44%; 45%
  - Overall coverage: 110x; 115x
  - ≥ 10X: 2,136,202; 2,128,057 bp (76.5%)
    - Custom base caller for males

X-linked disorder: Filtering

- Heterozygous
  - Carriers
- Severe
  - Non-synonymous, indels, nonsense, frameshifts
- Ultra-rare
  - Not in dbSNP, three concurrent controls

The Number of Genes with One or More Variants Following Each Filtering Criterion

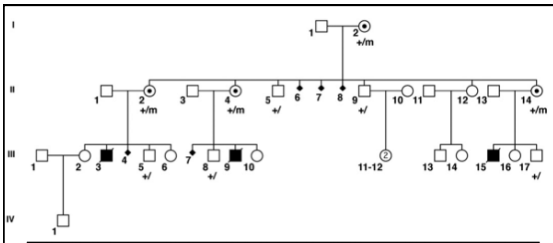
	All X Exons	
	Family 1	**Family 2**
Total substitutions	360	330
Heterozygous	271	229
Nonsynonymous	71	65
Not in dbSNP	14	16
Not in three controls	11	11
Nonsense	0	***1***
Total indels	53	47
Nonsynonymous	8	7
Not in dbSNP	3	2
Not in three controls	1	1
Frameshifting	***1***	0

X-linked disorder: Filtering

- An iterative process
  - Start stringent
  - Progressively relax
  - Minimizes variants to consider
- In this example, a 'hit' using first, most stringent filters: *RBM10*
- What if that was not the case?

The Number of Genes with One or More Variants Following Each Filtering Criterion

	All X Exons	
	Family 1	**Family 2**
Total substitutions	360	330
Heterozygous	271	229
Nonsynonymous	71	65
Not in dbSNP	14	16
Not in three controls	11	11
Nonsense	0	***1***
Total indels	53	47
Nonsynonymous	8	7
Not in dbSNP	3	2
Not in three controls	1	1
Frameshifting	***1***	0



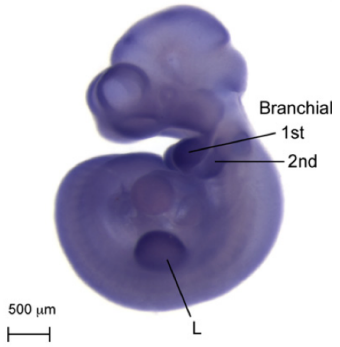
The Number of Genes with One or More Variants Following Each Filtering Criterion

	All X Exons		Linkage Region	
	Family 1	Family 2	Family 1	Family 2
Total substitutions	360	330	85	76
Heterozygous	271	229	54	54
Nonsynonymous	71	65	14	14
Not in dbSNP	14	16	5	4
Not in three controls	11	11	3	3
Nonsense	0	*1*	0	*1*
Total indels	53	47	9	7
Nonsynonymous	8	7	2	1
Not in dbSNP	3	2	1	0
Not in three controls	1	1	1	0
Frameshifting	*1*	0	*1*	0

X-Linked disorder: Filtering + linkage

- With linkage, variants drop a lot
- Difference from meiotic mapping
  - Any amount helps
  - LOD <3 OK
  - Be careful

Supportive evidence



- Two families with null mutations
- Absent in many controls
- Expressed in mouse in correct tissues
- Not strongest of evidence but type I error less likely with X

Johnston et al, Am J Hum Genet 2010 86:743-748

Autosomal recessive: CMAMMA

- Severe childhood onset, rare, metabolic acidosis
- Excluded all known causes
- Sequenced *single trio*

Autosomal recessive: CMAMMA

- Sequenced *single trio*

Filter	Number of variants
Initial variants	114,467
Quality (MPG≥10)	89,537
Compound heterozygous/ homozygous	7,864
Nonsynonymous/nonsense/ splice/frame shift	1,376
Not in dbSNP	301
Not homozygous in controls or MAF >10%	134
Candidate genes with two variants	12
*ACSF3, FAM63B, FAM154B, HLA-A*0226, LAMA2, LAMB4, LOC728138, MUC4, MUC17, OR10AD1, PLCH1, SBDS*	

Autosomal recessive: CMAMMA

- Seven unrelated affecteds for confirmation
  - WES vs. Sanger is question
- Whether & how to use dbSNP

Filter	Number of variants
Initial variants	114,467
Quality (MPG≥10)	89,537
Compound heterozygous/ homozygous	7,864
Nonsynonymous/nonsense/ splice/frame shift	1,376
Not in dbSNP	301
Not homozygous in controls or MAF >10%	134
Candidate genes with two variants	12
*ACSF3, FAM63B, FAM154B, HLA-A*0226, LAMA2, LAMB4, LOC728138, MUC4, MUC17, OR10AD1, PLCH1, SBDS*	

dbSNP

- Helpful and dangerous to use
- Repository of variation *irrespective* of the relationship of the variant to disease
- Individual variants may be pathologic
  - Variants found in disease gene identification studies or from clinical path labs
- Cohorts may be sourced from people with disease
  - DNAs from patients with cardiac rhythm disorders
  - Tedious to dig down to this level

dbSNP

- Your causative variant may be in dbSNP
  - As filtering is iterative, one may use it early on
  - For careful refinement, use MAF cutoffs
    - Try 5x-10x estimated frequency of disorder
    - CFTR example – 70% alleles delPhe508

dbSNP vs. other controls

- Consider using other sources
  - Your other exomes
    - Methodology match
  - 1000genomes, ClinSeq, et al
- Any can trip you up – again must set thoughtful thresholds and re-examine

Gene name	REF AA	VAR AA	AA POS	CDPred score	dbID	GENO-TYPES	HOM REF	HETS	HOM NONREF
ACSF3	L	P	2	-4	rs7188200(C,T)	179	40	87	52
ACSF3	R	W	10	0	-	464	463	1	0
ACSF3	A	P	17	-5	rs11547019(C,G)	499	452	47	0
ACSF3	G	S	64	-7	-	561	560	1	0
ACSF3	P	A	209	-12	-	290	289	1	0
ACSF3	P	L	285	-12	-	575	565	10	0
ACSF3	R	W	286	-10	-	575	574	1	0
ACSF3	R	L	318	-11	-	537	536	1	0
ACSF3	E	K	359	-9	-	574	573	1	0
ACSF3	V	M	372	-5	rs3743979(A,G)	564	33	232	299
ACSF3	R	Q	469	-6	-	572	567	5	0
ACSF3	R	W	471	-14	-	572	570	1	1
ACSF3	W	*	536	-30	-	555	554	1	0
ACSF3	R	W	558	-11	-	506	503	3	0

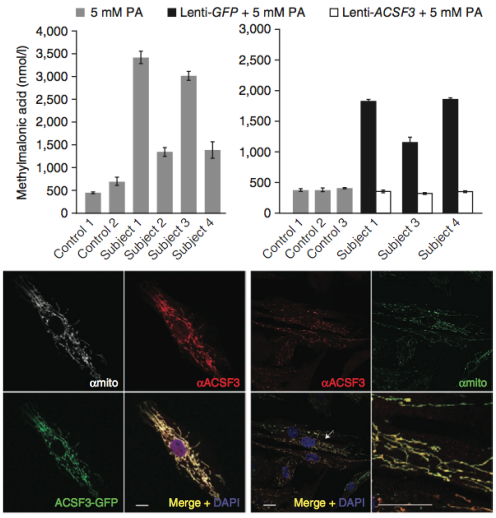
Gene name	REF AA	VAR AA	AA POS	CDPred score	dbID	GENO-TYPES	HOM REF	HETS	HOM NONREF
ACSF3	L	P	2	-4	rs7188200(C,T)	179	40	87	52
ACSF3	R	W	10	0	-	464	463	1	0
ACSF3	A	P	17	-5	rs11547019(C,G)	499	452	47	0
ACSF3	G	S	64	-7	-	561	560	1	0
ACSF3	P	A	209	-12	-	290	289	1	0
ACSF3	P	L	285	-12	-	575	565	10	0
ACSF3	R	W	286	-10	-	575	574	1	0
ACSF3	R	L	318	-11	-	537	536	1	0
ACSF3	E	K	359	-9	-	574	573	1	0
ACSF3	V	M	372	-5	rs3743979(A,G)	564	33	232	299
ACSF3	R	Q	469	-6	-	572	567	5	0
ACSF3	R	W	471	-14	-	572	570	1	1
ACSF3	W	*	536	-30	-	555	554	1	0
ACSF3	R	W	558	-11	-	506	503	3	0

Clinical evaluation

- 66 yo female
- Four accidents, poor memory, incontinence
- Serum and urine analysis
  - MMA plasma 48 uM (100x ULN), urine 70x ULN
  - MA plasma 11 uM (nl undetectable)

- Be careful – your controls may have your disease!

Supporting evidence



- 7/8 other patients with two mutations
- Dog with mutation
- Correction with transfected gene
- Localization
- Hypothesis-generating case

Sloan et al. Nat Genet 2011 43:883-886

Autosomal dominant: Hajdu-Cheney syndrome

- Very rare
- Dominant with many simplex
  - Progressive focal bone destruction
  - Characteristic radiographic abnormalities
  - Craniofacial anomalies
  - Renal cysts

Simpson et al. Nature Genetics 43, 303–305 (2011)
Isidore et al. Nature Genetics 43, 301–303 (2011)

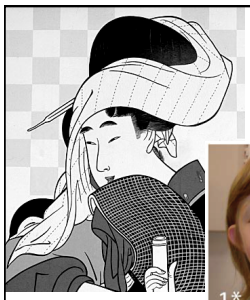
Sequencing & filtering approach

- Same as NISC
- 3-4 Gb / sample
- Two simplex and one multiplex family proband

- Filtering criteria:
  - Nonsynon, nonsense, splice or indel
  - *Never* before observed in dbSNP131, 1000G, or 40 controls
  - All three cases mutation in same *NOTCH2*

Follow-up, supportive evidence

- Sanger sequence 12 kindreds
  - 11 with mutations
  - Seven simplex – six with two parents confirmed as *de novo*
- *No functional data!*



Kabuki syndrome

- Dysmorphic, skeletal, immunologic, mild intellectual disability
- 1/30,000-1/50,000
- Most simplex, few vertical transmission

Ng et al, Nature Genet 42, 790–793 2010

Exome capture & sequencing

- Sequence ten unrelated exomes
  - Did not use trio – *de novo* strategy
- Somewhat different than current NISC approach
  - Selection by hybridization to custom exome arrays
  - ~6 Gb/patient, 40x coverage mappable regions

Original filter scheme

Any X of 10	1	2	3	4	5	6	7	8	9	10
NS/SS/ Indel	12,042	8,722	7,084	6,049	5,289	4,581	3,940	3,244	2,486	1,450
Not in dbSNP 1000G	7,419	2,697	1,057	488	288	192	128	88	60	34
Not in controls	7,827	2,865	1,025	399	184	90	50	22	7	2
Not in either	6,935	2,227	701	242	104	44	16*	6	3	1

- Asked how often a gene name appeared among their cases
- No good candidates identified
- "However, there was no obvious way to rank these candidate genes"

Filter strategy #2: Clinical stratification

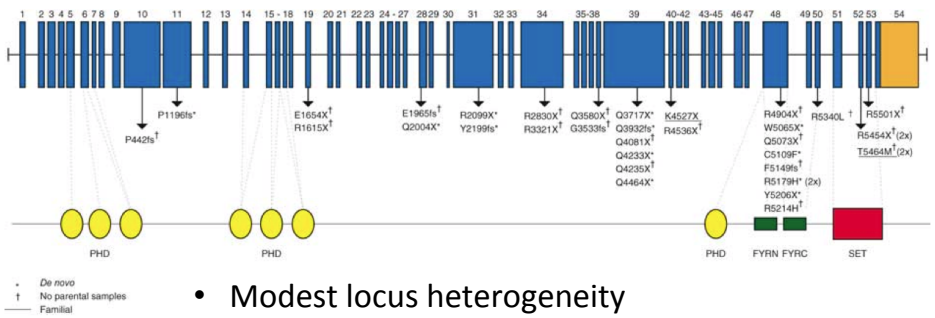
- Several clinicians ranked patients typical>atypical
- Predicted functional assessment of variants
- "Manual review of these data highlighted distinct, previously unidentified nonsense variants in *MLL2* in each of the four highest-ranked cases."
- Mutations in cases 1-4, 6, 7 & 9. No other gene with mutations in >2

Manual curation & follow-up genotyping

- 96% next gen coverage
- Sanger sequence *MLL2* in mutation-negative cases
  - Frameshift mutation missed in rank cases 8 & 10
- 43 additional cases > Sanger sequence
  - NonSyn, FS, & NS mutations in 26/43
- 12/12 cases with both parents were *de novo*

MLL2 gene mutations

Figure 1: Genomic structure and allelic spectrum of *MLL2* mutations that cause Kabuki syndrome.



- Modest locus heterogeneity
- Broad allelic heterogeneity
- Competing group announced they had failed - *MLL2* was not targeted by their capture!
- No functional data – early paper

Proteus syndrome

- Asymmetric overgrowth
- Nevi in lines of Blaschko
- Vascular malformations

- Never familial
- Discordant monozygotic twins

Happle Model: A somatic gene mutation, lethal in non-mosaic state

Explains:

Mosaic lesions

Absence of uniform cases

Absence of recurrences

Discordant monozygotic twins

Happle Model: A somatic gene mutation, lethal in non-mosaic state

Exome Sequence:

Four affected-unaffected sample pairs (n=8)

Two affected patients (n=3)

Parents (n=5)

Unaffected Monozygotic twin (n=1)

Happle Model: A somatic gene mutation, lethal in non-mosaic state

Exome Sequence Sample Types:

- Skin biopsy cultures
  - From clinically affected/unaffected areas
- Surgical specimens
  - Harvested in OR with clinical researcher in attendance

Happle Model: A somatic gene mutation, lethal in non-mosaic state

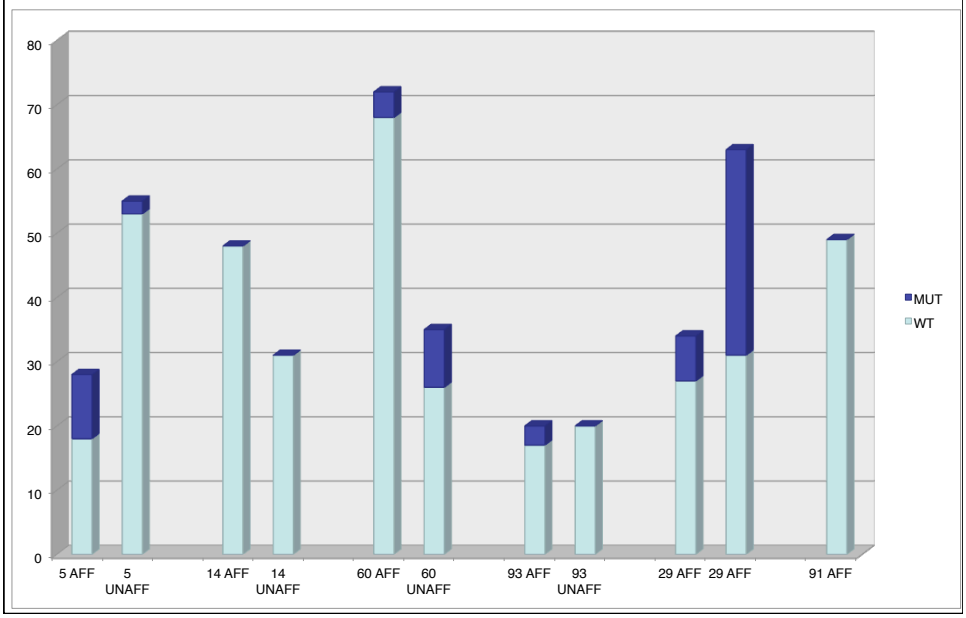
Exome Sequence Sample Types:

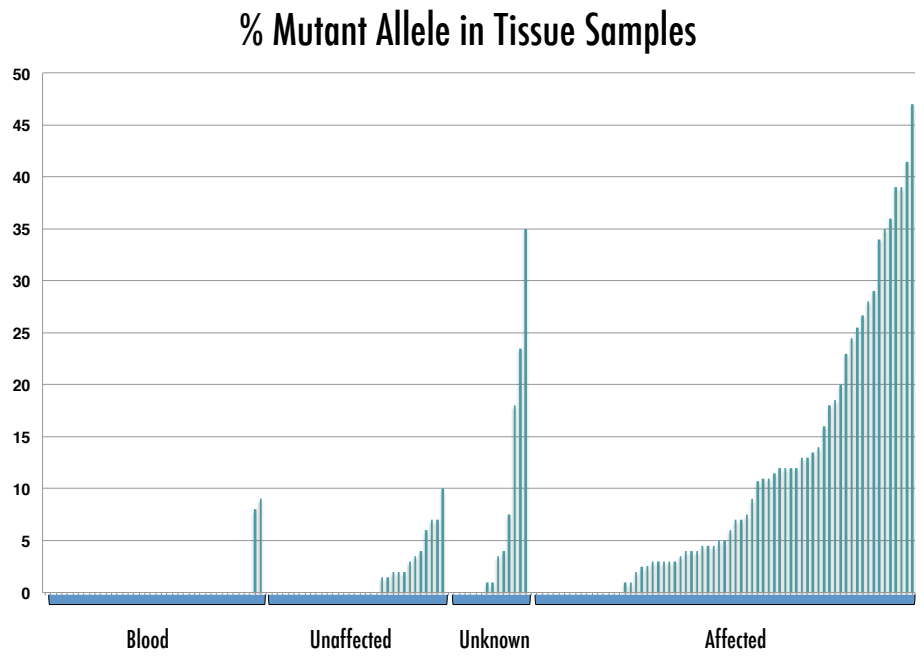
- Skin biopsy cultures
  - From clinically affected/unaffected areas
- Surgical specimens
  - Harvested in OR with clinical researcher in attendance
- *Did not use blood cell DNA*
  - *No hematopoietic phenotype*

Filtering criteria

- Nonsynon, NS, splice, indel
- Absent in dbSNP
- 100 – 300 differences in many of the pairs
- Validated with Sanger
- One persisted

Exome Data: g.chr14:104,317,596C>T

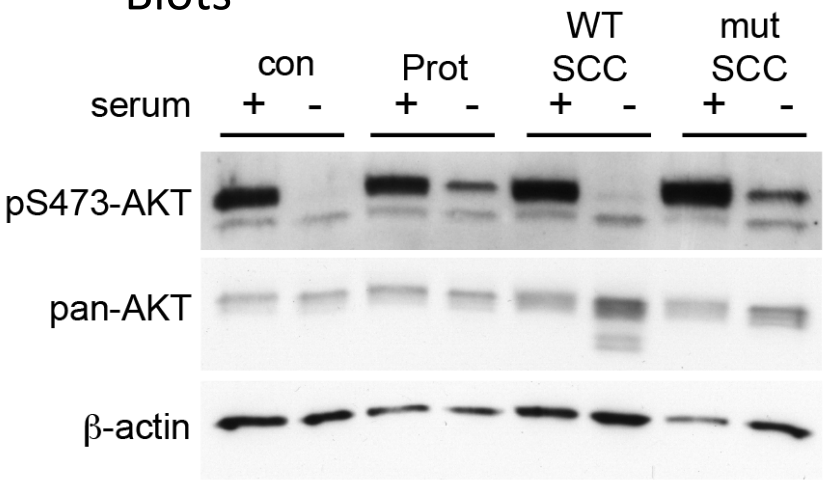




Summary of Mutation Survey

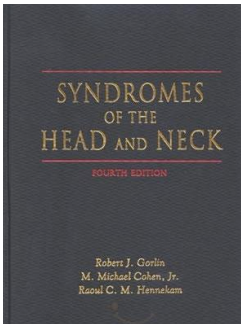
- 29/31 patients have identical g.chr14:104,317,596C>T
  - Two patients w/o mutation clinically similar
- Mutation more often found in grossly affected tissues
- Mutation rare in peripheral blood
- Not found in controls
  - ClinSeq (572 exomes): 0 sequence reads
  - 1000 genomes: 1 sequence read in ~30,000 (0 calls)

αpAKT Western Blots



Implications

- Deluge of disease mutation IDs
  - Syndromes Head & Neck
    - >2,500 entities
  - London Medical Database
    - >4,500 entities
  - Few with genes w known function, natural history, or management
  - Challenge both clinical and basic science

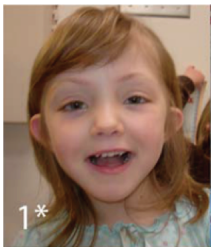


LONDON MEDICAL DATABASES

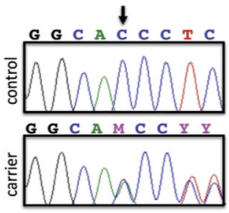
Welcome to London Medical Databases website

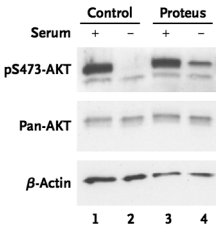
Implications II

- Exome or WGS will likely become a useful clinical diagnostic tool
- Algorithms and approaches developed in research will diffuse out into practice



The *whole* picture





Thanks to…

- J Johnston
- J Sapp
- F Facio
- J Teer
- Many trainees & staff
- NIH Intramural Sequencing Center
- Venditti group
- Stacie Loftus
- Andy Baxevanis
- Dave Kanney